INFORMATION AND CONTROL 11, 561–567 (1968)

An Analog of a Theorem about Context-Free Languages

BARRON BRAINERD

Department of Mathematics, University of Toronto, Toronto, Canada

INTRODUCTION

The well-known theorem of Bar-Hillel et al. (Bar-Hillel, 1964) to the effect that for every infinite context-free grammar G there exist integers p and q such that for each word α in $L(G)$, the language generated by G, if α contains more than p terminal symbols, then

(i) α can be written $\alpha = \beta_1\gamma_1\delta\gamma_2\beta_2$ where $\gamma_1\gamma_2$ is a non-null string and $\gamma_1\gamma_2$ contains fewer than q terminal symbols, and

(ii) $\beta_1\gamma_1{}^n\delta\gamma_2{}^n\beta_2$ belongs to $L(G)$ for all $n \geqq 1$

has the following corollary:

COROLLARY (I). *If L is an infinite language generated by a context-free grammar, then L contains a sequence, $\{\alpha_n\}$, of strings such that the sequence of lengths $\{\ell(\alpha_n)\}$ is a (nontrivial) arithmetic progression.*

The converse of this corollary is not true, as can easily be seen by considering the language $L_1 = \{a^n b^n a^n \mid n \geqq 1\}$, which has been shown to be non-context-free in Bar-Hillel (1964). That (I) does not hold in general for languages generated by context-sensitive grammars follows from the following example:

EXAMPLE 1. (Due to J. Friant, 1966). Consider the context-sensitive grammar G_2 with nonterminal vocabulary S, $\#$, A, terminal vocabulary a, initial string $\#S\#$, and rules

$$\#S \to \#A,$$

$$AS \to S^2A,$$

$$A\# \to S^2\#,$$

$$S \to a.$$

This grammar generates the language

$$L(G_2) = \{a^{2^n} \mid n \geqq 0\}.$$

Since $L(G_2)$ contains no (infinite) sequences which form arithmetic progressions, (I) does not hold for context-sensitive languages.

The purpose of this note is to extend (I) to a class of languages that meets, but perhaps does not contain, the context-free languages. This class is a subclass of the languages generated by matrix grammars developed by S. Abraham (1965). As a corollary to our main result, we find that there are context-sensitive grammars which do not fall into our subclass.

In this note we employ, for the most part, the standard nomenclature and notation of the mathematical theory of generative grammars. See either S. Ginsburg (1966) or S.-Y. Kuroda (1964).

1. MATRIX GRAMMARS

If V is a finite set of symbols (the vocabulary), then $F(V)$ denotes the set of all finite nonempty strings of elements of V. A *language over* V is a subset of $F(V)$. For any $\beta \varepsilon F(V)$, $\ell(\beta)$ denotes the number of symbols in β, sometimes called the *length of* β. The symbol Λ denotes the empty string. Note that $\Lambda \varepsilon F(V)$ and $\ell(\Lambda) = 0$.

Let $G = \langle V_N, V_T, S, \mathbf{R}\rangle$ be a context-free grammar in the following sense: V_N is a finite set of symbols (the *nonterminal vocabulary*), V_T is a second finite set of symbols disjoint from V_N (the *terminal vocabulary*), S is an element of V_N (the *initial symbol* of G), and $\mathbf{R}$ is a finite set of rules (the *production rules* of G) which take the form $A \to \omega$ where $A \varepsilon V_N$, and $\omega \varepsilon F(V_N \cup V_T)$ (i.e., ω is a nonempty string of symbols from $V_N \cup V_T$).

An ordered set

$$M = [R_1, R_2, \cdots, R_m] \tag{1}$$

for $m \geqq 1$ and $R_i \varepsilon \mathbf{R}$ $(i = 1, 2, \cdots, m)$ is called a *matrix of rules over* G. A finite set $\mathbf{M}$ of such matrices is called a *matrix grammar* over G.

If $R = (A \to \omega)$ is a rule of $\mathbf{R}$, and $\alpha \varepsilon F(V_N \cup V_T)$ has the form $\alpha = \alpha_1 A \alpha_2$, then we write $R(\alpha) = \alpha_1 \omega \alpha_2$. It should be stressed that the correspondence which carries α into $R(\alpha)$ is, in general, a many-valued function from some subset (the *domain of* R) of $F(V_N \cup V_T)$ into $F(V_N \cup V_T)$. Sometimes we signify the relation $\beta = R(\alpha)$ by writing

$$\alpha \xrightarrow{R} \beta,$$

which can be read α *goes into* β *under the production rule* R. If M is the matrix in (1), we write $\beta = M(\alpha)$, provided there exist $\alpha_1, \alpha_2, \cdots,$

α_{m-1} in $F(V_N \cup V_T)$ such that

$$\alpha \xrightarrow{R_1} \alpha_1 \xrightarrow{R_2} \alpha_2 \xrightarrow{R_3} \cdots \xrightarrow{R_m} \alpha_m = \beta.$$

Sometimes we write the relation $\beta = M(\alpha)$ as

$$\alpha \overset{M}{\Rightarrow} \beta$$

and say β is *generated* from α by means of M. If there exist $M_1, M_2, \cdots, M_k \in \mathbf{M}$ such that

$$\alpha \overset{M_1}{\Rightarrow} \beta_1 \overset{M_2}{\Rightarrow} \beta_2 \overset{M_3}{\Rightarrow} \cdots \overset{M_k}{\Rightarrow} \beta_k = \beta, \tag{2}$$

we say β *is generated from* α and write $\alpha \Rightarrow \beta$. The *language generated* by the matrix grammar $\mathbf{M}$ is then the set

$$L(\mathbf{M}) = \{\beta \in F(V_T) \mid S \Rightarrow \beta\}. \tag{3}$$

The sequence $\alpha, M_1, \beta_1, M_2, \beta_2, M_3, \cdots, M_k, \beta$ of (2) is called a *derivation of* β *from* α.

To illustrate these concepts, we construct a matrix grammar $\mathbf{M}_1$ for the language $L_1 = \{a^n b^n a^n \mid n \geqq 1\}$. The underlying CF-grammar of $\mathbf{M}_1$ is $G_1 = \langle\{S, X, Y, Z\}, \{a, b\}, S, \mathbf{R}_1\rangle$ where $\mathbf{R}_1$ contains the rules

$$S \to XYZ, \qquad X \to a(X), \qquad Y \to b(Y), \qquad Z \to a(Z).$$

The matrices of $\mathbf{M}_1$ are

$$M = [S \to XYZ],$$
$$M' = [X \to aX, Y \to bY, Z \to aZ],$$
$$M'' = [X \to a, Y \to b, Z \to a].$$

To generate $a^n b^n a^n$, first apply M, then apply M' $n - 1$ times and finally apply M'' to obtain

$$S \overset{M}{\Rightarrow} XYZ \overset{M'}{\Rightarrow} aXbYaZ \overset{M'}{\Rightarrow} \cdots \overset{M'}{\Rightarrow} a^{n-1}Xb^{n-1}Ya^{n-1}Z \overset{M''}{\Rightarrow} a^n b^n a^n. \tag{3}$$

It is clear that the only derivations possible with $\mathbf{M}_1$ are of the form of (3). Thus

$$L(\mathbf{M}_1) = L_1.$$

Since an arbitrary CF-grammar is also a matrix grammar (where all the matrices have length 1), this example yields the result that matrix grammars have more generating power than CF-grammars.

2. INDEX OF A MATRIX GRAMMAR

Let $\mathbf{M}$ be a matrix grammar over the CF-grammar $G = \langle V_N, V_T, S, \mathrm{R}\rangle$. For $\alpha \varepsilon F(V_N \cup V_T)$, let $d(\alpha)$ stand for the string formed by deleting all terminal symbols from α.

Thus, if $\alpha = aAbCDeFg$, where $A, C, D, F \varepsilon V_N$ and $a, b, e, g \varepsilon V_T$, then $d(\alpha) = ACDF$. Note that the condition that α is in the domain of a matrix M depends only on the form of $d(\alpha)$ and not on the number or the configuration of the terminal symbols in α.

If for some $k \geqq 1$ and some $\beta \varepsilon F(V_T)$

$$D : S, M_1, \sigma_1, M_2, \sigma_2, \cdots M_k, \sigma_k = \beta$$

is a derivation of β from S, then the *index* of the derivation D is

$$i(D) = \max_{j=1,2,\cdots k} \ell(d(\sigma_j)).$$

Clearly the index of a derivation of $\beta \varepsilon F(V_T)$ from S is always a non-negative integer. Let D_β be the set of all derivations of $\beta \varepsilon L(\mathbf{M})$ from S. With every string $\beta \varepsilon L(\mathbf{M})$ we associate a number

$$i(\beta) = \min_{D \varepsilon D_\beta} i(D),$$

the *index* of β. A derivation D of $\beta \varepsilon L(\mathbf{M})$ from S is called *admissible* if $i(D) = i(\beta)$.

A matrix grammar $\mathbf{M}$ is of *finite index* if there is a natural number N such that $i(\beta) \leqq N$ for all $\beta \varepsilon L(\mathbf{M})$. The smallest possible value of N is the *index* of $\mathbf{M}$. Clearly the matrix grammar M_1 of the previous section has index 3.

To provide a grammar with nonfinite index, consider the following matrix grammar $\mathbf{M}_3$ over the CF-grammar:

$$G_3 = \langle \{S, A, B\}, \{a, b\}, S, \mathrm{R}_3\rangle,$$

where R_3 contains the rules

$$S \to \begin{Bmatrix} AS \\ B \end{Bmatrix}, \qquad B \to \begin{Bmatrix} B \\ b \end{Bmatrix}, \qquad A \to a.$$

The matrices of $\mathbf{M}_3$ are

$$[S \to AS], \qquad [S \to B], \qquad [B \to b],$$

and

$$[A \to a, B \to B].$$

Clearly $L(\mathbf{M}_3) = \{a^n b \mid n \geqq 1\}$ and $i(a^n b) = n + 1$ for each $n \geqq 1$. Therefore $\mathbf{M}_3$ does not have finite index.

3. THE MAIN RESULT

To prove the generalization of (I) mentioned in the introduction, we need the following lemma:

LEMMA 1. *If* $\mathbf{M}$ *is a matrix grammar of finite index say* m_0, *then there exist a finite subset* Q *of* $F(V_N) \cup \{\Lambda\}$ *such that each* $\beta \varepsilon L(\mathbf{M})$ *has an admissible derivation.*

$$\sigma_0 = S, M_1, \sigma_1, M_2, \sigma_2, \cdots, M_k, \sigma_k = \beta,$$

where $d(\sigma_j) \varepsilon Q$ *for all* $j = 0, 1, \cdots, k - 1$.

Proof. Obviously, Q is a subset of $F_{m_0}(V_N) \cup \{\Lambda\}$, where

$$F_{m_0}(V_N) = \{\alpha \varepsilon F(V_N) \mid \ell(\alpha) \leqq m_0\}.$$

Now we state and prove the result indicated in the introduction.

THEOREM 1. *If* $\mathbf{M}$ *is a matrix grammar of finite index, then either* $L(\mathbf{M})$ *is finite or it contains an infinite sequence* $\{\beta_n\}$ *such that* $\{\ell(\beta_n)\}$ *forms an arithmetic progression.*

Proof. Let the cardinality of the set Q associated with $\mathbf{M}$ by Lemma 1 be k. Assume $L(\mathbf{M})$ is not finite. Then there is a (sufficiently long) string $\sigma \varepsilon L(\mathbf{M})$ with an admissible derivation D involving the application of $q > k$ matrices that extend the lengths of the strings in their domains. Suppose the derivation D is given by

$$\sigma = \mu_{q+1} M_q \mu_q \cdots M_1 \mu_1 (S),$$

or alternatively by

$$S \overset{\mu_1}{\Rightarrow} \alpha_1 \overset{M_1}{\Rightarrow} \beta_1 \overset{\mu_2}{\Rightarrow} \alpha_2 \overset{M_2}{\Rightarrow} \cdots \overset{\mu_q}{\Rightarrow} \alpha_q \overset{M_q}{\Rightarrow} \beta_q \overset{\mu_{q+1}}{\Rightarrow} \sigma,$$

where for each i, μ_i is a possibly empty string of matrices each of which does not extend the length of strings upon which it can act, and for each i, M_i is a matrix which does extend the length of the strings upon which it can act. Since $q > k$, there exist natural numbers j, ℓ such that

$$1 \leqq j < \ell \leqq q$$

and

$$d(\beta_j) = d(\beta_\ell).$$

If $\nu = M_\ell \mu_\ell \cdots M_{j+1} \mu_{j+1}$, then

$$\beta_\ell = \nu(\beta_j);$$

and since $d(\beta_j) = d(\beta_t)$, ν can be applied to β_t to yield $\nu(\beta_t) = \nu^2(\beta_j)$ with $d(\nu^2(\beta_j)) = d(\beta_j)$. This process can be iterated an arbitrary number of times to yield $\nu^n(\beta_j)$ with

$$d(\nu^n(\beta_j)) = d(\beta_j) = d(\beta_t).$$

Therefore, for each $n \geqq 1$, we can apply

$$\mu_{q+1}M_q\mu_q \cdots M_{t+1}\mu_{t+1}$$

to $\nu^n(\beta_j)$ to obtain a terminal string $\beta_n \varepsilon L(\mathbf{M})$; i.e.,

$$\beta_n = \mu_{q+1}M_q\mu_q \cdots M_{t+1}\mu_{t+1}\nu^n(\beta_j).$$

Each application of ν introduces a fixed positive number, say r, of new terminal symbols, so that the number of terminal symbols in $\nu^n(\beta_j)$ is

$$nr + \ell(\beta_j) - \ell(d(\beta_j)).$$

Since the composition

$$\mu_{q+1}M_q\mu_q \cdots M_{f+1}\mu_{f+1}$$

of matrices introduce a fixed number s of terminal symbols,

$$\ell(\beta_n) = \{s + \ell(\beta_j) - \ell(d(\beta_j))\} + nr$$

for $n \geqq 1$ and s, r fixed natural numbers. Hence $\{\ell(\beta_n)\}$ is an arithmetic progression.

4. COROLLARY AND UNSOLVED PROBLEMS

COROLLARY 1. *The* CS-*language* $L_2 = \{a^{2^n} \mid n \geqq 0\}$ *does not have a matrix grammar of finite index.*

Proof. It has already been observed in Example 1 that L_2 contains no sequence $\{\beta_n\}$ such that $\{\ell(\beta_n)\}$ is an arithmetic progression. Therefore by Theorem 1, L_2 cannot be generated by a matrix grammar of finite index.

Consideration of the following pair of open questions may prove of interest:

A. Does every CF-grammar have finite index?[1]

B. A matrix grammar can be viewed as a CF-grammar where the class of allowable derivations is restricted by the matrix requirement. The matrix requirement, however, is not the only possible type of

[1] If this conjecture is valid, then Theorem 1 is a generalization of (I).

restriction on derivations. Can Theorem 1 be proved for CF-grammars when other kinds of restrictions are placed on the allowable derivations?

RECEIVED: April 14, 1967

REFERENCES

ABRAHAM, S. (1965), Some questions of phrase structure grammars I. *Computational Ling.* **4**, 61–70.

BAR-HILLEL, Y. (1964), "Language and Information." Addison-Wesley, Reading, Massachusetts, 116–150. Also Bar-Hillel et al. under the title: On formal properties of simple phrase structure grammars, in *Z. Phonetik, Sprachwiss. Kommunikationforschung* **14** (1961), 143–172.

FRIANT, J. (1966), "Les Langages CS." Thèse présentée à la Faculté des Sciences de l'Université de Paris.

GINSBURG, S. (1966), "The Mathematical Theory of Context-free Languages." McGraw-Hill, New York.

KURODA, S.-Y. (1964), Classes of languages and linear-bounded automata. *Inform. Control* **7**, 207–223.